Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-84-AC1, 2016

Hydrology and
Earth System
Sciences
Discussions
Open Access

Interactive comment on “A statistically based seasonal precipitation forecast model with automatic predictor selection and its application to Central and South Asian headwater catchments” *by* Lars Gerlitz et al.

Lars Gerlitz et al.

lars.gerlitz@gfz-potsdam.de

Received and published: 3 June 2016

Dear referee,

thank you very much for your comments concerning our manuscript and particularly for the detailed remarks on the utilized statistical techniques.

Please find enclosed our response as well as some suggestions which will hopefully improve the presented manuscript.

Best regards, Lars Gerlitz et al.

##

1) I only have some training in statistics, and don’t have education background in hydrology or climate. My viewpoint may be quite different from the HESS community. Technically, the forecast model is well-designed using various domain knowledge. I would like to know how k-means is carried out, e.g., whether the correlation is used in the clustering, how to determine the random seeds or initial cluster centers. As we know, different random seeds will lead to different clustering results in K-means.

The cluster analysis is performed using the Algorithm of Hartigan & Wong (1979), which is generally known as reliable and computationally efficient. The seeds are set randomly and are recursively updated afterwards in order to minimize the sum of squares between the observations and their assigned cluster center. In theory, the approach might be slightly sensitive to the choice of seeds, however, in practice we did not experience any variations of the clustering solution. In the revised manuscript, we will give some advanced information concerning the clustering techniques.

Hartigan, J. A. and M. A. Wong (1979) : A k-means clustering algorithm". Applied Statistics28.1, pp. 100-108.

2) Simply aggregating neighboring grids to a big region does not make sense if these grids have quite different correlation with the precipitation in the target region.

As stated in our manuscript, the clusters are not constructed based on spatial distances. The clustering routine is applied to the normalized time series of each potential predictor grid cell. This approach identifies clusters, which are characterized by a similar temporal variability of the considered predictor variable. Thus it is secured, that all grid cells within one cluster have a similar correlation with the precipitation time series.

3) How can simply aggregating ’poor’ monthly forecast to seasonal forecast lead to some good forecast results?

Monthly precipitation amounts in general are characterized by a large (random) noise, which is due to non-predictable meso- or local scale circulation patterns. The random noise often results in large magnitudes of variability on shorter time scales. On a seasonal scale, random events are most likely averaged out and the observed precipitation time series is less influenced by single events. Thus the hindcast based on large scale predictors can better reproduce the observations. We will clearly point that out in the revised manuscript.

4) Before using random forest, we should explore a single regression tree for monthly forecast first in order to identify which kinds of predictors are more important as well as its performance.

Due to the complex structure of the model and the high number of (partially highly correlated) predictor variables, we believe, that the investigation of single regression trees does not lead to a better interpretability of the model results. Single regression trees could only be constructed for one particular month and lead time, which would result in an overall number of 12*6=72 regression trees. The fact, that many of the predictor variables are correlated (e.g. the ENSO-variables and all of its covariates), impedes the interpretation of single regression trees.

5) Reasonable forecast comparison should be carried out and reported. For example, what is difference if only those widely accepted climate indexes are used as predictors. How about comparing random forecast with a single regression tree. How about using 7 clusters instead of 7?

The manuscript proposes one forecast methodology which especially stands out due to the negligence of traditional climate indices and the automatic predictor selection. Our main aim is to show, that a forecast based on data mining techniques without any prior knowledge is feasible. This has been shown by means of four case studies. We believe, that a comparison with different kinds of statistical models (traditional vs. automatic predictors / random forest vs. regression trees) would go beyond the scope of the manuscript.

Concerning the number of clusters, we absolutely agree, that this is a very subjective decision (which however might be a useful adjustment screw). As stated in the manuscript, “an excessive number of clusters might result in a disjunction of predictor regions, which reduces the predictive skill. On the contrary an insufficient number of clusters will lead to an aggregation of large regions which might still be characterized by a large inhomogeneity and thus are not suitable for the derivation of potential predictor variables“ (p7l12). Results (not shown) indicate, that a low number of clusters leads to an underestimation of variability, while a large number of clusters rather leads to random like results. For future activities we will definitively think about automated solutions and will discuss this issue in the “discussion” part of the presented manuscript. Further we suggest to highlight the model sensitivity to the number of clusters in the “methods” section but resign from a comparison of different model results with variable cluster numbers. The latter would result in a confusing manuscript structure and would not create additional information.

6) The sensitive analysis to me does not fit the key statement of paper that much. Actually from the appearance frequency of predictors in random forest can give us some ideas how important is a predictor.

Unfortunately the calculation of simple predictor importance measures is only reliable for uncorrelated predictor variables. In general the random forest importance parameter for one particular predictor variable is based on the increase of the model error, which results from the modification of this variable (permutation importance). If there are highly correlated variables in the predictor space, every variable can easily be substituted, which results in unrealistically low values of the random forest importance measure

see e.g.: Gregurutti et al., 2014,: Correlation and variable importance in random forests, Statistics and Computing, DOI: 10.1007/s11222-016-9646-1).

Furthermore the aggregation to seasonal precipitation forecasts leads to a black boxlike model structure and impedes the interpretability. Thus we believe, that the sensitivity / response analysis based on traditional climate indexes is a good possibility to test the plausibility of the model. Since the results agree with former studies on the inter-annual precipitation variability in Central and South Asia, we argue that the sensitivity analysis gives evidence, that the model is able to find important variables (among others, those, which represent well known atmospheric modes) and to quantify their contribution. In a revised version we will try to better communicate the reasons for and the results of the sensitivity analysis.

7) Line 42, Page 15 has some typo or overclaim. AUC values are not always >0.7.

Will be changed.

8) There are too many references. Reference Chen et al (2012) needs some correction.

We will try to shorten the introductory section in some parts and correct the mentioned reference!

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-84, 2016.